Testing the randomness of causality networks from multivariate time series

Dimitris Chorozoglou[1] and Dimitris Kugiumtzis[2]

[1]Department of Mathematics, Aristotle University of Thessaloniki, Greece,
Email: dchorozo@hotmail.com
[2]Department of Electrical and Computer Engineering,
Aristotle University of Thessaloniki, Greece
Email: dkugiu@auth.gr

Abstract—In network analysis it is important to contrast the given or formed network to a random network, typically by means of significance testing of some network measures. This requires the generation of random networks that preserve certain properties of the original network, i.e. the total number of nodes and connections or even the number of connections of each node. In this paper we show that these schemes are not appropriate for correlation or causality networks formed from multivariate time series, which have nodes the observed variables and connections given by some measure of correlation (undirected connections) or Granger causality (directed connections). Further, we propose a scheme that performs the randomization on the time series rather than the network connections. Particularly for the networks formed by cross correlation, we generate surrogates for each time series preserving the marginal distribution and linear autocorrelation, and form the network from the surrogate multivariate time series. Simulations on multivariate time series with no inter-dependencies shows that the classical network randomization erroneously tends to reject the null hypothesis of random network, whereas the proposed scheme does not.

1. Introduction

In recent years, networks have been used in the analysis of multivariate time series, which have as nodes the observed variables and their connections are determined by a correlation-based index for non-directed connections [1, 2], or a causality-based index for directed connections [3]. For the correlation networks in particular, the correlation coefficient is used to give the weighted connections or its significance to give binary connections, where the significance is determined either by thresholding or a hypothesis test for significance.

In network analysis, a question of interest in many settings is whether the network is random. To address this question, first random networks are generated by randomizing the connections preserving the total number of connections (total degree) or total sum of connection weights (total strength), or even the degree distribution [4, 5, 6]. Then the original network is compared to these networks using appropriate network measures [1]. In this work, we show that these schemes for randomization of networks are not appropriate for correlation networks from multivariate time series, and we propose an appropriate method for this by randomizing the time series. We show the superiority of the proposed method on time series independent to each other for different settings of autocorrelation, time series length and number of variables.

2. Randomization of networks

The problem of interest in this work is to generate random networks with the same number of nodes as the given network and random connections. In the generation of random networks certain features are preserved. The simplest random network generation is simply by randomly shuffling the original connections, so that the total degree for binary connections or the total strength for weighted connections is preserved [1]. A more stringent condition is to preserve the degree or strength of connections of each node. Among different schemes, we consider here the following scheme of Maslov and Sneppen [7] for binary connections. For two randomly selected connections (i, j) and (k, l) of the original network, the end nodes j and l are interchanged giving rise to the connections (i, l) and (k, j). If these connections are not already in the network they are created and the first two connections are deleted, otherwise the step is repeated. After many iterations, this process creates a randomized variant of the original network that preserves the number of connections at each node. We simply refer to this method as Maslov.

3. Correlation networks

In correlation networks, nodes correspond to random variables and the connections are given by an association measure, such as the Pearson correlation coefficient [2]. We assume that for a set of K random variables $\{X_{1,t}, \ldots, X_{K,t}\}$, a sample of n time ordered observations for each variable are given, $\{x_{k,1}, \ldots, x_{k,n}\}$ for $k = 1, \ldots, K$, which can be completely independent, dependent within each variable (having autocorrelation), and dependent across variables (having cross-correlation). As the interest in this study is on undirected network connections, we consider as the association measure the linear zero-lag cross-correlation, which is actually the Pearson correlation

coefficient. For two variables $X = X_i$ and $Y = X_j$, $i, j \in \{1, \ldots, K\}$, the coefficient is defined as $r_{X,Y} = s_{XY}/\sqrt{s_X^2 s_Y^2}$, where $s_{XY} = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})$ is the sample covariance of (X, Y), $s_X^2 = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})^2$ is the sample variance of X, and $\bar{x}$ is the sample mean of X. Further, the cross-correlation $r_{X,Y}$ is converted to a valid weight connection of X and Y nodes, $w_{X,Y}$, and here following [8] we set $w_{X,Y} = b_{X,Y}/b_{\max}$, where $b_{X,Y} = r_{X,Y} - \bar{r}_{X,Y} + 1$, $\bar{r}_{X,Y}$ is the average of $r_{X,Y}$ and $b_{\max}$ is the maximum of $b_{X,Y}$ for all pairs (X, Y).

It is known that the correlation matrix R having at each entry (i, j) the Pearson correlation coefficient of X_i and X_j, $r_{i,j}$, is positive semi-definite. This holds regardless of the correlation structure of the random variables, and thus also when the network is completely random. For binary connections, the adjacency matrix A of zeros and ones is derived from R by assigning one only when the correlation is found significant by some criterion. The criterion can be an arbitrary threshold or a threshold corresponding to a specified density of connections, or resulting from statistical significance test [2]. The condition of positive semi-definiteness of R determines accordingly conditions for the feasible forms of W and A.

The network randomization methods in Sec. 2 are initially developed for networks given by the studied system, and there are no conditions on the node connections, as is the case for the correlation networks. Thus the resulting correlation network by such methods may be more random than a random correlation network derived from the respective uncorrelated variables. For correlation networks from time series, the effect of the autocorrelation on the cross-correlation and thus the formation of connections should also be addressed.

4. Randomization of networks by randomizing the multivariate times series

We propose a method to generate random networks addressing the constraints on the correlation matrix R, and subsequently the weight matrix W and the adjacency matrix A. The rationale is to randomize the time series rather than the connections (a similar approach was first implemented in [9]). Each of the K time series $\{X_{1,t}, \ldots, X_{K,t}\}$ for $t = 1, \ldots, n$, is randomized separately under the condition of preserving the marginal distribution and the autocorrelation function [10], or equivalent the power spectrum [11]. Here, we use the algorithm of Iterative Amplitude Adjusted Fourier Transform (IAAFT) in [11] because for very small time series considered in the study it provides the least variance in the match of autocorrelation. The procedure forming B randomized correlation networks with weighted connections is given in the following steps:

1. For each time series $\{X_{k,t}\}_{t=1}^{n}$, $k = 1, \ldots, K$, a randomized (surrogate) time series $\{X_{k,t}^*\}_{t=1}^{n}$ is generated by IAAFT, giving the surrogate multivariate time series $\{X_{1,t}^*, \ldots, X_{K,t}^*\}_{t=1}^{n}$.
2. The correlation matrix R^* is computed on $\{X_{1,t}^*, \ldots, X_{K,t}^*\}_{t=1}^{n}$.
3. A proper weight matrix W^* is formed by R^*.
4. Steps 1-3 are repeated B times to generate B randomized correlation networks.

The time series $\{X_{1,t}^*, \ldots, X_{K,t}^*\}_{t=1}^{n}$ are by construction independent to each other, and the entries of the correlation matrix R^*, and subsequently the weights in W^*, are all insignificant corresponding to independent variables. It is known that strong autocorrelation may give rise to spurious cross-correlation, and therefore cross-correlation entries in R^* may be found statistically significant using a parametric significance test. This implies that a strong connection in the original network may correspond to such spurious cross-correlation and then this will be preserved in the randomized network as the particular pair of surrogate time series preserves the original autocorrelations.

For unweighted networks, an adjacency matrix A^* replaces the weight matrix W^* in step 3, and two approaches for forming A^* are considered. First, the same threshold criterion (a given threshold value or significance test) is applied as for the original network, which however does not preserve the total degree. To preserve the total degree, the threshold that gives the total degree of the original network is used.

Having generated B randomized networks with any method, we test the null hypothesis H_0 that the original network is random, i.e. the observed variables are uncorrelated and any connections formed are random, using a proper network measure as test statistic. Here, we consider the following network measures (their notations in parentheses): clustering coefficient (ClustCoeff), betweenness centrality (BetwCentr), eigenvector centrality (EigenCentr), characteristic path (CharPath), global efficiency (GlobEffic), average degree or strength (Degr/Strn), assortativity (Assortat), density (Density), eccentricity (Eccentric), diameter (Diameter), all defined for weighted and unweighted networks[1]. Each network measure q is computed on the original network giving the value q_0, and B randomized networks giving the values $q_1, \ldots, q_B$. The p-value of the test is $2r_0/(B+1)$ if $r_0 < (B+1)/2$ and $2(1-r_0)/(B+1)$ if $r_0 > (B+1)/2$, where r_0 is the rank of q_0 in the ordered list of $q_0, q_1, \ldots, q_B$.

5. Simulation study and results

The simulation study is on independent to each other time series as the interest is on specificity, i.e. whether a true random network is actually found not to be different

[1] We used the Matlab functions of the Brain Connectivity Toolbox in https://sites.google.com/site/bctnet/measures

from randomized networks generated by each method, and thus the H_0 that the original network is random is not rejected. First we give an example of an unweighted network formed by 15 independent time series of length $n = 200$, all with strong autocorrelation, and the threshold 0.3 for the cross-correlation was used to determine the binary connections (see Fig. 1g). For a pair of two original time series

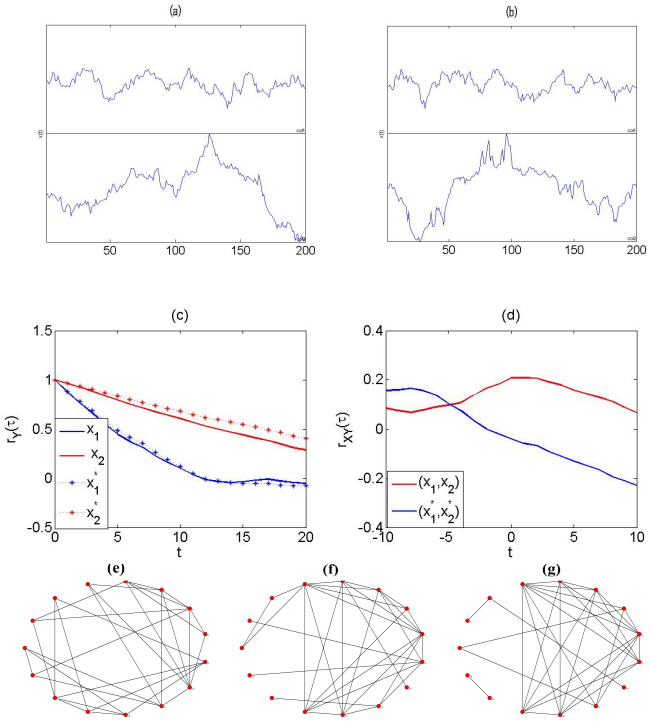

Figure 1: Two of the fifteen original time series in (a) and the corresponding surrogate time series in (b), their autocorrelations in (c) and their cross-correlation in (d). The randomized network with the proposed method in (e) and the Maslov method in (f) for the original network in (g).

shown in Fig. 1a, one pair of surrogate time series generated by IAAFT is shown in Fig. 1b, and their autocorrelation and cross-correlation as functions of the lag are shown in Fig. 1c and d, respectively. Though the autocorrelation and cross-correlation of the surrogate time series do not match accurately these of the original time series, there are no systematic or large differences. The unweighted network using the threshold that gives the same number of connections as for the original network is shown in Fig. 1e and the network generated by the Maslov method is shown in Fig. 1f. One may find differences by eyeball judgement in the original network in Fig. 1g and the two randomized networks in Fig. 1e and f, but to obtain statistically valid results we make Monte Carlo simulations and perform the significance test of random network for the different network measure statistics and the different methods generating randomized networks.

In the simulations, we applied the three variants of our proposed approach, denoted RTSweight for weighted connections, RTSbinthr using a given threshold for drawing binary connections, and RTSbindeg using a threshold for drawing the same number of connections as for the original network. We compared these methods to two methods generating random networks, RNavestr that preserves the average (or total) strength for weighted networks, and RNnoddeg that preserves the node degree for unweighted network (method of Maslov).

Independent autocorrelated time series were generated by first order autoregressive processes, AR(1), where the strength of autocorrelation was given by the coefficient ϕ of AR(1). Different settings were used for varying ϕ, number of time series K and time series length n. For each setting, 100 multivariate time series (Monte Carlo realizations) were generated and for each realization $B = 100$ randomized networks were formed by each of the five methods, and thereafter 10 network measures were computed on each network. Using each network measure as test statistic, the corresponding p-value was calculated and the test decision was reached at the significance level 0.05. The percentage of rejections of H_0 in the 100 realizations are reported in Fig. 2 in the form of color maps (gray scale, black for zero and white for 100%) for network measure vs method for random network generation, and for each of the six data settings. The two methods shuffling ran-

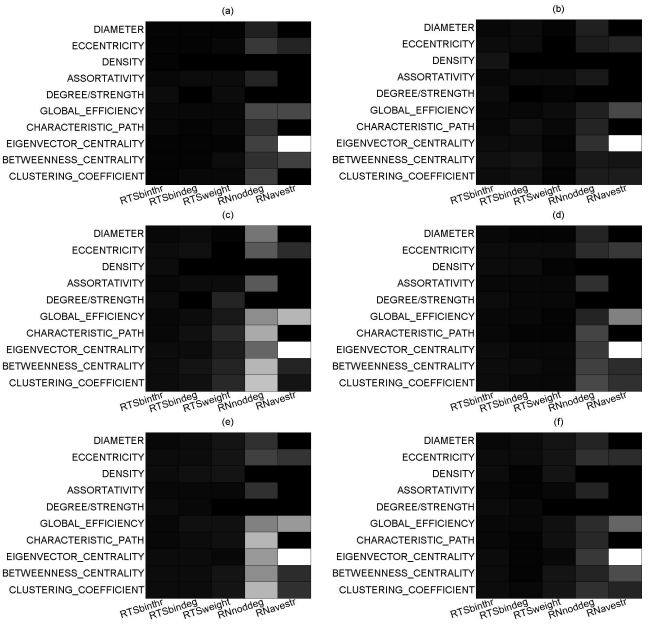

Figure 2: Color map for the estimated probability of rejection of H_0 of random network on 100 realizations of a different setting at each panel, for 10 network measures (row) and 5 methods for random network generation (column). (a) $K = 15$, $n = 200$, $\phi = 0.9$. (b) $K = 15$, $n = 200$, $\phi = 0.6$. (c) $K = 15$, $n = 50$, $\phi = 0.9$. (d) $K = 15$, $n = 50$, $\phi = 0.6$ (e) $K = 25$, $n = 200$, $\phi = 0.9$ (f) $K = 10$, $n = 50$, $\phi = 0.6$.

domly the connection, RNavestr and RNnoddeg, tend to reject H_0 with higher probability than the nominal probability of 0.05 (high type I error), which varies with the network measure. This is observed for all data settings and the type I error seems to increase with the autocorrelation strength (two last columns in all colored maps). Method RNavestr fails particularly when eigenvector centrality is used as test statistic, a measure related to the property of semi-definiteness of the correlation matrix. On the contrary, all three variants of the proposed approach of randomizing the time series give rejection rates at the nominal significance level, indicating proper specificity obtained by these methods. This is also derived from the summary results for all network measures for each data setting in Table 1 [2].

Table 1: Average rejection rate of H_0 over all network measures for each method (row) and the six data settings (column) termed from (a) to (f) as in Fig. 2.

Method	(a)	(b)	(c)	(d)	(e)	(f)
RTSbinthr	4.1	6.1	5.4	5	5.3	5.4
RTSbindeg	4	7.2	7.6	4.3	6.1	3.5
RTSweight	6	2.6	13.1	3	8.2	8.6
RNnoddeg	34.2	12.6	75.1	29.7	71.6	16.5
RNavestr	47.8	37.8	51.8	39.2	52	25.6

6. Conclusion

The results of the simulation study highlighted two main points of the work: a) the insufficiency of network randomization methods applied directly to the correlation network in generating networks that have the same characteristics as the original network when the original network is indeed derived from uncoupled variables, b) the effectiveness of the proposed approach randomizing the time series rather than the network connections in generating random networks sharing the same characteristics as the original network derived from uncoupled variables. The proposed approach was developed for both unweighted and weighted networks and for correlation thresholds set arbitrarily to derive binary connections or aiming to preserve the same number of binary connections as the original network. The preservation of network characteristics was assessed by ten network measures used as test statistics for a randomization test of significance, i.e. assessing whether the network measure of the original network falls within the distribution of the network measure computed on the randomized networks. The simulations on different settings of autocorrelation strengths, number of variables and time series length showed that the proposed approach performed equally well at the different data settings. The data settings considered in this study aimed at assessing the specificity of the proposed approach for the generation of randomized networks. Next, we plan to test the sensitivity by repeating the same simulation setup on coupled stochastic processes, e.g. vector autoregressive processes.

[2] From the average rejection rate over network measures, measures that reflect the conditions preserved by a method such as the average strength or degree were exempted for this method.

Acknowledgement

DK is supported by the Greek General Secretariat for Research and Technology (Aristeia II, No 4822).

References

[1] M. Newman. *Networks: An Introduction*. Oxford University Press, New York, 2010.

[2] S. Horvath. *Weighted Network Analysis, Applications in Genomics and Systems Biology*. Springer, New York, 2011.

[3] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon. Econometric measures of connectedness and systemic risk in the finance and insurance sectors. *Journal of Financial Economics*, 104(3):535–559, 2012.

[4] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[5] M. E. J. Newman, S. H. Strogatz, and D. J. Watts. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 64:026118, 2001.

[6] C. I. Del Genio, H. Kim, Z. Toroczkai, and K. E. Bassler. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLoS ONE*, 5(4):E10012, 2010.

[7] S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296(5569):910–913, 2002.

[8] G. Ansmann and K. Lehnertz. Surrogate-assisted analysis of weighted functional brain networks. *Journal of Neuroscience Methods*, 208(2):165–172, 2012.

[9] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. The backbone of the climate network export. *EPL (Europhysics Letters)*, 87(4):48007, 2009.

[10] D. Kugiumtzis. Statically transformed autoregressive process and surrogate data test for nonlinearity. *Physical Review E*, 66:025201, 2002.

[11] T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77(4):635–638, 1996.